Language is a primary means through which stereotypes and prejudice are communicated and perpetuated.

(Hamilton and Trolier, 1986; Bar-Tal et al., 2013)

Social bias in crowdsourced datasets

- Crowdsourcing annotations has become a fundamental aspect of NLP research.
- What are the ethical implications of soliciting crowdsourced data, specifically social biases that may emerge when asking for generated sentences.

Handson session

- Perform a “bias audit” of an NLP dataset produced by crowdsourcing.
- You will attempt to measure the presence of social stereotypes in this dataset that may have harmful effects if used to train classifiers in downstream tasks.
- After this analysis, you will present specific examples from the data that you speculate could be particularly biased and problematic.

Credits: 11-830 Computational Ethics in NLP. Maarten Sap

Tools

- You will use **pointwise mutual information (PMI)** to find which associations are being made with identity labels.
- PMI can be used as a measure of word association in a corpus, i.e. how frequently two words co-occur above what might just be expected based on their frequencies.
- Here we use PMI to measure which words co-occur with labels for identities. This allows us to see associations that may perpetuate stereotypes.

Data and Resources

- **SNLI corpus**, a popular dataset for the NLP task of natural language inference. You can read more about the dataset and task on the SNLI website or in the original paper (Bowman et al. 2015).
- Use the training data CSV or JSON lines file (**snli_1.0_train.jsonl**) in the SNLI corpus. The sentence1 column with premise that was supplied to annotators, and the sentence2 column is the hypotheses that annotators came up with. See other details about the corpus format in the README supplied with SNLI.
- **List of identity labels** (based on Rudinger et al. 2017).

Word association analysis

- Function should take *a unigram, with word frequencies* relative to a corpus, as input and give a list of other *unigrams in the corpus ranked by PMI*.
- Terms that occur less than 10 times in the corpus should not be considered; optionally you can consider other thresholds.
- For preprocessing, lowercase, remove stopwords and tokenize the data.
- Note that there are duplicate premises and hypotheses in the data; remove these and just look at unique utterances.

PMI

- $c(wi)$ be the count of word wi in the corpus
- c(wi,wj) be the number of times that wi and wj occur in the same premise or hypothesis.
- N as the number of documents (premises or hypotheses)
- P(wi) as the word frequency c(wi)/N.
- Then PMI is:

$$PMI(w_i, w_j) = \log_2 \frac{P(w_i, w_j)}{P(w_i)P(w_j)} = \log_2 \frac{N \cdot c(w_i, w_j)}{c(w_i)c(w_j)}$$

- Compute PMI between the identity labels in the provided list and all other words in the SNLI training corpus
- Look at the top associated words for identity labels of your choice.
- Do you see any that may reflect social stereotypes?
- It is helpful to compare the top PMI words for certain identity terms with other related ones (such as men compared with women).
- Note that some terms in the list do not occur in the data; they are included for advanced analysis on possible other corpora